R E P O R T R E S U M E S

ED 020 239 UD 005 673

ISSUES IN EVALUATING PROGRAMS FOR DISADVANTAGED CHILDREN.

BY- FOX, DAVID J.

PUB DATE DEC 67

EDRS PRICE MF-$0.25 HC-$0.20 3P.

DESCRIPTORS- *PROGRAM EVALUATION, *DISADVANTAGED YOUTH, *EDUCATIONAL PROGRAMS, *RESEARCH PROBLEMS, *EDUCATIONAL RESEARCH, EVALUATION METHODS, EVALUATION CRITERIA, EVALUATION NEEDS, PROGRAM EFFECTIVENESS, RESEARCH UTILIZATION, EDUCATIONAL RESEARCHERS,

EVALUATING PROGRAMS FOR THE DISADVANTAGED IS DIFFICULT BECAUSE THERE IS NO CLEAR THEORETICAL OR OPERATIONAL DEFINITION OF THE TERM "DISADVANTAGED." THE PROLIFERATION OF SCHOOL PROGRAMS ALSO CREATES EVALUATION PROBLEMS, FOR OFTEN EVALUATION TAKES PLACE BEFORE THE PROGRAM IS EITHER FULLY IMPLEMENTED OR HAS FUNCTIONED FOR A SUFFICIENT PERIOD OF TIME. A DISTINCTION SHOULD BE MADE BETWEEN CRITICAL EVALUATION, WHICH WOULD DETERMINE THE CONTINUATION OF A PROGRAM, AND ONGOING EVALUATION, WHICH WOULD IDENTIFY ITS STRENGTHS AND WEAKNESSES. RESEARCHERS MUST RECOGNIZE THAT THOSE WITH VESTED INTERESTS IN THE SCHOOL POWER STRUCTURE WILL USE EVALUATIONS TO SUPPORT THEIR OWN VIEWPOINTS. EVALUATIONS SHOULD ESTABLISH SENSIBLE, RELEVANT, MEANINGFUL CRITERIA OF SUCCESS OR FAILURE AND RECOGNIZE THE TOTALITY OF GAINS AND LOSSES OF A GIVEN ENDEAVOR. THERE IS ALSO THE QUESTION OF WHAT IS TO BE CONSIDERED PROGRESS FOR CHILDREN WITH DEFICITS--WHETHER IT SHOULD BE DEFINED AS RESTORATION TO THEIR NORMAL BEGINNING LEVEL, OR AS ARRESTING ANY FURTHER DECLINE, OR AS RAISING THE LEVEL TO THAT OF NONDEFICIENT CHILDREN. ALSO, RESEARCHERS CAN NO LONGER MAINTAIN AN ISOLATION FROM POTENTIAL IMPLEMENTATION OF THEIR FINDINGS BUT MUST BECOME INVOLVED IN THE USES TO WHICH THEIR DATA ARE PUT. THIS ARTICLE WAS PUBLISHED IN "THE URBAN REVIEW," VOLUME 2, NUMBER 3, DECEMBER 1967. (NH)

U D005 673

ED 020 239

Testing and Evaluation / Two Views

However one characterizes American education today, there can be no denying the fundamental importance of evaluation *—of measuring the effectiveness (or simply the effect) of the thousands of educational programs, practices and strategies that are now, as never before, under the public scrutiny.*

For, whether the trend of the times is toward confusion, indecision and futility or toward dynamic and thoroughgoing innovation —or both at the same time —our primary need is to know in detail what is actually going on *in the classrooms and school systems of the nation.*

This need predicates another, the design and specification of accurate, reliable and sensitive systems of observation, measurement, testing and, in the final analysis, judgment.

But since there are few human transactions so subtle, so infinitely variable and fast-moving, as the process of education, the construction of devices to discover and communicate what was or is happening to those concerned is an extraordinarily difficult undertaking.

At present, some of these devices, so Messers. Hunter and Rogers point out, seem woefully crude. They are a bit like an elephant trying to pick up a pea, as H. G. Wells once remarked in an (obviously) different connection. The point, of course, is not whether an elephant can *pick up a pea (he can) but whether so much apparent power is well-spent in the effort. Mr. Fox, in his discussion, reminds us that in assessing the worth of educational programs and practices we must shun the trivial as well as the irrelevant and be severely rigorous in matching the* expected *outcomes of a particular endeavor with what, if anything, came out of it.*

This is not to say there is no interest in such a fact as that certain "programs" have enhanced teacher morale without noticeably improving the reading ability of their pupils. But it is to say that if these programs were primarily intended to improve the childrens' reading, then a very serious question has been raised and the architects of the program must either revise their goals to include morale building or set about other means of [illegible] or old) educational of these practices mselves, it is con- tting the burden American schooling.

ED020239

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Testing: Politics and Pretense

by Lloyd B. Hunter and Frederick A. Rogers

Within the last decade or so, standardized tests have been accorded equal, if not greater, importance than the traditional three R's themselves. In the face of increasing changes of emphases in educational practice and methodology, they are being looked to as the one objective standard of measurement that can tell us whether our pupils are making adequate gains in achievement. Certainly, every parent has come to know (be it through personal experience or vicarious exposure) the importance of standardized tests for his child. Performance on these tests, particularly in the early grades, has often played a large role in determining whether the child will be promoted, and if promoted, where he will be placed within the grade level. By the same token, school administrators have sought to use the results of these tests as aids in assessing the effectiveness of educational programs, and in deciding *where* within their systems to put *what* types of educational resources and services.

Issues in Evaluating Programs for Disadvantaged Children

by David J. Fox

05613

The *Handbook of Research on Teaching* is a 1172 page compilation of research prepared in the early sixties under the auspices of the American Educational Research Association. In the extremely detailed index, which covers 47 pages, there are only three references to "evaluation" and the material referred to covers six of the 1172 pages. This is some indication of the little attention paid to serious evaluation in education in the years covered by the handbook. More recently however, because of the congressional mandate to "evaluate" the programs for disadvantaged children established under the Elementary and Secondary Education Act, we have had a flood of evaluative studies. Since many of these studies are presumably being conducted to guide decisions as to the continuation, expansion or elimination of the programs, it seems appropriate to take a critical look at some of the issues involved in evaluating programs for disadvantaged children.

In a sense, the ambiguity of the terms defines the issues. Taking them in

Within the last decade or so, standardized tests have been accorded equal, if not greater, importance than the traditional three R's themselves. In the face of increasing changes of emphases in educational practice and methodology, they are being looked to as the one objective standard of measurement that can tell us whether our pupils are making adequate gains in achievement. Certainly, every parent has come to know (be it through personal experience or vicarious exposure) the importance of standardized tests for his child. Performance on these tests, particularly in the early grades, has often played a large role in determining whether the child will be promoted, and if promoted, where he will be placed within the grade level. By the same token, school administrators have sought to use the results of these tests as aids in assessing the effectiveness of educational programs, and in deciding *where* within their systems to put *what* types of educational resources and services. On a national level, too, the federal government's dramatically increased involvement in education has resulted in an increase of programs for particular groups of pupils, and the effectiveness of these programs will, in most instances, be judged, again, on the basis of standardized tests.

It is the contention of this paper that both the design of these tests and the ways in which their results are currently reported render them inadequate as tools for assessing the overall value of educational programs. This is especially true for those programs designed specifically for those pupils whose performances may be described as occupying either extreme of a given distribution of scores.

Unquestionably, most public attention in recent years has centered on the groups of pupils whose performances are at the lower end of their distribution. This, in turn, has resulted in a marked proliferation of proposals and programs for the improvement of pupil performance and achievement. However, the construction of appropriate and reliable measures of pupil performance has not kept pace with the introduction of "new" educational practices. Educators today are still applying basically the same criteria and standardized instruments that were arrived at almost a decade ago.

We have no intention here of judging the value of particular programs. Our position is simply that standardized tests are too insensitive to be considered effective guides to the kind of evaluation that should lead to needed innovation and desirable changes. To see why this is so, we shall have to take a close look at how these tests are standardized.

Continued on page 6

The *Handbook of Research on Teaching* is a 1172 page compilation of research prepared in the early sixties under the auspices of the American Educational Research Association. In the extremely detailed index, which covers 47 pages, there are only three references to "evaluation" and the material referred to covers six of the 1172 pages. This is some indication of the little attention paid to serious evaluation in education in the years covered by the handbook. More recently however, because of the congressional mandate to "evaluate" the programs for disadvantaged children established under the Elementary and Secondary Education Act, we have had a flood of evaluative studies. Since many of these studies are presumably being conducted to guide decisions as to the continuation, expansion or elimination of the programs, it seems appropriate to take a critical look at some of the issues involved in evaluating programs for disadvantaged children.

In a sense, the ambiguity of the terms defines the issues. Taking them in inverse order, consider first the word "disadvantaged." We have never defined the term sufficiently for research or evaluation purposes. We began at a point where it was cultural deprivation, but soon recognized that this was an absurdity in that all children have a culture and none is truly deprived of culture. We moved on to terms like "socially disadvantaged," "educationally disadvantaged," and now it has become simply "disadvantaged," which, of course, returns us to the question: disadvantaged in terms of what? At some level, one might say a group is "disadvantaged" because every researcher in the country is studying it and every curriculum specialist is developing a special curriculum for it and the disadvantage is that the group is so busy being corrected, remedied, tested, experimented with, and having their aspiration level raised, there is relatively little time to learn. More seriously, however, researchers doing program evaluation should have insisted for themselves (just as they do for their students) on clearly defined and clearly specified theoretical and operational definitions of the subject of their research. And these definitions should have been agreed upon *before* doing research or developing programs or systems for evaluation.

By leaving the term "disadvantaged" so vague, social scientists have encouraged such a broad application of the label that the public might be forgiven for coming to the conclusion that every school with any significant proportion of Negro or Puerto Rican pupils is a school for disadvantaged

Continued on page 7

children, or that there are no *white* disadvantaged children, or that there are no Negro and Puerto Rican *advantaged* children. The effect of these simplistic overgeneralizations on the larger problems of intergroup relations is a question that merits serious study. More relevant here, however, is their effect on the public's perception of the public school.

One possible effect is that public education has come to be seen by people of all ethnic backgrounds as virtually synonymous with compensatory education. And a possible consequence of this is that the parent who doesn't perceive his children to be in need of special help will exert every effort and spend whatever is necessary to remove his offspring from the public schools and place them in schools dedicated to providing "normal" or even "enriching" instruction. Indeed, I believe that this is one of the most important factors contributing to the withdrawal of middle-class allegiance from the public educational system. If one does not see oneself as socially or economically "oppressed," or one's children as particularly different from other kids of his age in the society, why on earth would one risk incurring the stigma of sending those children to "special" schools for the disadvantaged?

The second term I should like to look at is *program*. Obviously, to evaluate an educational program you must have a program. But what is that? Traditionally, the usual educational program is thought of as a long-planned, carefully developed set of materials and/or practices which might be implemented on a system-wide basis after successful tryout on a smaller scale. But at present almost any idea anyone has for which he can find some financing immediately becomes a program. One has only to consider the proliferation of those "programs" in the public schools that provided a massive introduction of "program" specialists in art, music, remedial reading, corrective reading, and corrective math. One wonders, not altogether facetiously, how children who spend so much time in the specialized enrichment, remedial and corrective areas will have enough time in regular class to develop the deficits and retardation by which all these special programs are justified. We seem to believe that a program springs full blown into existence in the same way that Minerva is presumed to have come into being. Is there no gap between idea and implementation? No period of reflection? Is there no period of development in which a program should be left alone to get on its feet?

This suggests that to appraise a program with any reasonable hope of accuracy we should build into the design a developmental period in which evaluation is held in abeyance until the program is established. We should avoid premature evaluation at points in time when the program is so new that finding it is the major research problem. We evaluate at a point when researchers have acted as if there were only one kind, critical evaluation, as if the world were waiting for the outcome of our studies to decide what to do. It seems to me astonishing that in 1967 researchers have been so naive and so unaware of the politics of education as to believe that evaluation will have this type of effect. After all, these same researchers have been concerned for years over the lack of impact that research generally has had on educational policy and practice. Why do they suppose that evaluation would fare much better? If anything, evaluation, with all its connotations of an outsider checking up on the effectiveness of professional behavior, could be expected to widen the gap between practitioner and researcher which has so often been noted in the education literature. As Benjamin Bloom concludes in his chapter in the *Handbook of Research on Teaching*, "... the research worker must not expect major modifications of teaching practice in a brief period of time." There is no reason for the evaluation researcher to expect more than this.

Moreover, if researchers are to play an intelligent role in improving American education, they must understand the forces that determine the need for evaluation and the extent to which, and the ways in which, the results of evaluations will be used. It is obvious that every group involved in the power structure of the educational enterprise will seek to use research and evaluation data to support their own points of view. The fact that interest in educational research is so new may, in part, explain why researchers have not fully understood the extent to which their data can be used in this way. But if we are to play an effective role in the process, we must acquire this understanding with dispatch.

The next term I would like to consider is implicit in any consideration of evaluation; I mean the *criteria* by which we determine the success or failure of a program. My first problem here is to wonder why researchers have accepted, and even participated in the movement to abandon criteria which have served the schools and the process of education for generations. I refer to criteria like intellectual functioning as reflected in intelligence tests or traditional areas of academic functioning such as reading and arithmetic. It is ironic in New York City that *parents* have been responsible recently for the reintroduction and reemphasis on the basic academic criteria of progress in reading and arithmetic, after researchers and school people had willingly abandoned these in the interests of the vague, diffused, and difficult-to-measure criteria involving social and personal functioning. An impartial observer of recent years would conclude that these criteria had been abandoned and reinstated more for sociopolitical reasons than educational ones, but this only highlights the importance of researchers understanding the

immediately becomes a program. One has only to consider the proliferation of those "programs" in the public schools that provided a massive introduction of "program" specialists in art, music, remedial reading, corrective reading, and corrective math. One wonders, not altogether facetiously, how children who spend so much time in the specialized enrichment, remedial and corrective areas will have enough time in regular class to develop the deficits and retardation by which all these special programs are justified. We seem to believe that a program springs full blown into existence in the same way that Minerva is presumed to have come into being. Is there no gap between idea and implementation? No period of reflection? Is there no period of development in which a program should be left alone to get on its feet?

This suggests that to appraise a program with any reasonable hope of accuracy we should build into the design a developmental period in which evaluation is held in abeyance until the program is established. We should avoid premature evaluation at points in time when the program is so new that finding it is the major research problem. We evaluate at a point when the people attempting to implement the program are still learning about it, are insufficiently oriented, when materials required to implement it do not yet exist in the schools, and when the children have been assigned to it for no more than a session or two.

For example, the Head Start program, which moved public education into an entirely new area, which involved the training of vast numbers of personnel for new teaching and aide roles and which required materials that were in process of development as the program got under way—this program was nevertheless "evaluated" in its first summer!

Why have not researchers demanded adherence to a sensible work pace involving tryouts and testing with revision, followed by broad implementation only when there is clear evidence that the program is something that helps children and is a justifiable expenditure of public funds?

Allied to this is the confusion over the third term, *evaluation*. There are two major ways in which we can enter the public school system to evaluate. One is what might be called *critical evaluation*. That is an evaluation which will determine whether a program is to continue or is to be more widely implemented or is to be phased out. A second way is one that might be called *ongoing evaluation*, in which the question of continuing the program is not at issue, but where the evaluative function is to identify the strengths and weaknesses of the elements in the program with a view toward improving it. Not only have we not distinguished between these two kinds of evaluation, but ing with dispatch.

The next term I would like to consider is implicit in any consideration of evaluation; I mean the *criteria* by which we determine the success or failure of a program. My first problem here is to wonder why researchers have accepted, and even participated in the movement to abandon criteria which have served the schools and the process of education for generations. I refer to criteria like intellectual functioning as reflected in intelligence tests or traditional areas of academic functioning such as reading and arithmetic. It is ironic in New York City that *parents* have been responsible recently for the reintroduction and reemphasis on the basic academic criteria of progress in reading and arithmetic, after researchers and school people had willingly abandoned these in the interests of the vague, diffused, and difficult-to-measure criteria involving social and personal functioning. An impartial observer of recent years would conclude that these criteria had been abandoned and reinstated more for sociopolitical reasons than educational ones, but this only highlights the importance of researchers understanding the social context in which they function today.

A second aspect of criteria that troubles me is our willingness to use and suggest criteria that we know to be absurd. Improved school attendance appears as a criterion in almost every type of Title I project, as does improved self-image, and aspiration. Thus it seems as if we believe that any program, whether cultural, social, personal, or academically oriented, must have an effect on the students' self-image, aspiration, and attendance. But what if their self-images go up to the point that they decide they no longer need school? Attendance then should go down! But more important, we are setting impossible goals for ourselves and, in a sense, dooming a program to fail when we apply criteria that simple logic and the understanding of the process of education indicate it cannot achieve.

In establishing criteria we have also ignored two other basic issues. One is their *relative* importance when several criteria are sensible. For example, the study done for the Center for Urban Education in the spring of 1966, on the Free Choice-Open Enrollment Program in New York City revealed considerable evidence of social and personal gains for the children being bussed. There was also evidence that although the children being bussed gained in reading skills, they did not gain at the same rate as children in the sending schools who had remained there. Now, how do we combine these findings to

Continued on page 9

make some overall statement about the success of the program? Did the social gains compensate for the somewhat slower rate of improvement in reading or was the reading loss more important and significant enough to override the social gains? This problem of weighing criteria must be settled if an overall evaluation of a program is ever to be made.

We have also ignored what might be called the over-flow criteria in programs for the disadvantaged. For example, if we remove children from a regular class for a corrective reading program, what impact does this have on those who stay in class? What is the impact on resident children of bussing other children into their school? What is the impact of bussing on those who stay in the school from which other children are bussed? In selecting criteria we have concentrated on the children directly participating in the program, and we have ignored the effects on those who are not in the direct focus of the program.

My final concern with criteria is with the extent to which researchers have accepted certain assumptions and even stereotyped notions about children for which there is little substantial evidence. For example, what substantial evidence is there that the self-image of disadvantaged children is ubiquitously poor? What solid evidence is there that universally their aspirations levels are in need of improvement? Where are there normative data on self-image or aspirational level against which we can measure or test these assumptions? And if they do not exist, as indeed I believe they do not, why have researchers not concentrated some energy on obtaining these data?

In short, why hasn't the researcher been a voice of sense and sanity in the effort to define meaningful criteria capable of measurement and relevant to the program being evaluated?

Another term which is implicit in the effort to evaluate and which also needs defining is the term *progress*. If many theoreticians are right and children of disadvantaged populations, however defined, either enter school with a deficit or acquire a deficit within a year or two after entering school, then what is progress for such children? If we applied to the educational process Cannon's homeostatic thesis, which the physician uses as a definition of progress in treating physical illness, progress would be movement toward the restoration of his normal functioning. But what is normal functioning for a child who enters school with an initial deficit? Figure 1 illustrates two different ways of looking at this.

Figure 1

Conceptualization of Different Definitions of Progress

to his normal beginning level. But in the lower part of Figure 1 we see a child who enters with a deficit and who, soon thereafter, falls even further behind. If we intervene for him is it progress if we follow path A and the decline we would expect to continue is stopped? Is it progress if we follow path B and he is restored to the level of his original deficit? Is it progress if we follow path C and he is raised to what would be considered normal for other children? Or is it not progress until, as some parents are now insisting, he is raised to levels beyond what is a normal for the mass of children in school? In short, what is our own level of aspiration for the disadvantaged? In New York City it seems at times as if this level of aspiration is simply to bring everybody up to grade level so that we can eliminate the concern with retardation. This, of course, ignores the fact that for many children grade level is too low a level of aspiration and for many others it is too high.

Another aspect of the problem implied in this discussion is the researcher's role in the implementation of his findings. It seems to me that to do evaluation without a clear commitment to implementation raises a danger for the researcher that he will become a sociopolitical dupe used to give critics of programs a feeling that evaluation is being done. If ongoing evaluation is to be meaningful it must have provisions for immediate feedback into a program and the few instances of what I have called critical evaluation are even more in need of provision for implementation.

I do not believe that researchers can maintain their traditional isolation from implementation by arguing that their function is to evaluate in an objective way, leaving to others the responsibility for implementation. We are working in such complex areas with such difficult problems of data interpretation that we must begin to insist upon the right to participate in the decision-making process when it involves the interpretation and the application of our own findings. We must recognize that we are studying an issue about which people are concerned. We have finally become social scientists in a vivid sense of the term. Since our problems and our data now have social, economic, and political implications, I feel that the intelligent researcher must insist upon being involved in the use of these data.

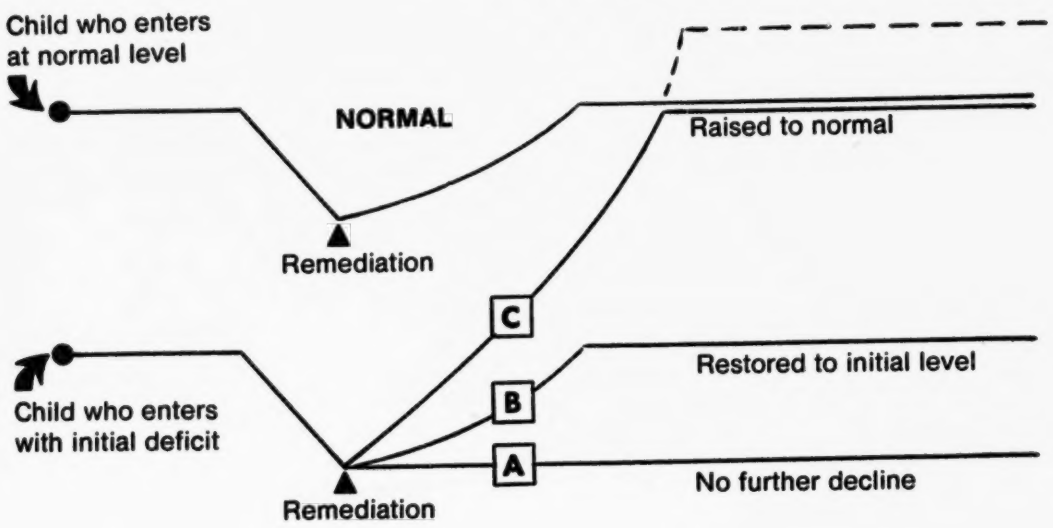
Finally, the last issue implicit in the discussion is the very nature of programs for the education of the disadvantaged. Where are the researchers who are willing to challenge the current orientation of such programs, namely, their exclusive preoccupation with efforts to correct, to supplement, to remediate, to compensate. Why have all our criteria been designed to test whether these corrective and remedial programs have worked? Why have we not sought criteria and encouraged programs concerned with building on the *strengths* of disadvantaged children? Scanning current efforts, one can only wonder if these children have any strengths. I think of my own recent experience in a [illegible] with a second grade teacher who was troubled by her disadvantaged

the program being evaluated?

Another term which is implicit in the effort to evaluate and which also needs defining is the term *progress*. If many theoreticians are right and children of disadvantaged populations, however defined, either enter school with a deficit or acquire a deficit within a year or two after entering school, then what is progress for such children? If we applied to the educational process Cannon's homeostatic thesis, which the physician uses as a definition of progress in treating physical illness, progress would be movement toward the restoration of his normal functioning. But what is normal functioning for a child who enters school with an initial deficit? Figure 1 illustrates two different ways of looking at this.

Figure 1

Conceptualization of Different Definitions of Progress



Consider a child who enters school at what we consider a normal beginning level. Then assume a problem arises and he drops below normal. We intervene with remediation and consider it successful if he is restored

working in such complex areas with such distinct p
tation that we must begin to insist upon the right to participate in the decision-making process when it involves the interpretation and the application of our own findings. We must recognize that we are studying an issue about which people are concerned. We have finally become social scientists in a vivid sense of the term. Since our problems and our data now have social, economic, and political implications, I feel that the intelligent researcher must insist upon being involved in the use of these data.

Finally, the last issue implicit in the discussion is the very nature of programs for the education of the disadvantaged. Where are the researchers who are willing to challenge the current orientation of such programs, namely, their exclusive preoccupation with efforts to correct, to supplement, to remediate, to compensate. Why have all our criteria been designed to test whether these corrective and remedial programs have worked? Why have we not sought criteria and encouraged programs concerned with building on the *strengths* of disadvantaged children? Scanning current efforts, one can only wonder if these children have any strengths. I think of my own recent experience in a discussion with a second-grade teacher who was troubled by her disadvantaged pupils who had difficulty reading, difficulty in arithmetic, and who she felt lacked sufficient self-control to maintain any atmosphere she considered proper for a second-grade class. As an example of the troubles she faced, she discussed a seven-year-old child who was often absent from school. Upon inquiry she discovered he preferred, on a school day, to go on the subway, sneak on a train, and go to Macy's to browse. It never dawned on her that, compared to some "advantaged" children who are unable to travel two blocks to school by themselves, this disadvantaged child had strengths she could use and build on. He had independence, responsibility, sophistication, freedom from fear, initiative and curiosity, to say nothing of experiences on which a program could be built. While not directly concerned with the issue of evaluation, I would like to see at least one program developed with this positive orientation and I would be delighted to face the new problems of evaluating its success.

David J. Fox is Director of the Office of Research and Evaluation Services at the School of Education of the City College of the City University of New York. He is also Chairman of the department of social and psychological foundations at the City College.